Conceptual Schemata for Terminology: A Continuum from Headings to Values in Patient Records and Messages

Angelo Rossi Mori, Elena Galeazzi, Fabrizio Consorti, W. Dean Bidgood, Jr., M.D., M.S.
Reparto Informatica Medica, ITBM-CNR, Roma, Italy
IV Clinica Chirurgica, Univ. La Sapienza, Roma, Italy
Duke University Medical Center, Durham, NC and National Library of Medicine, Bethesda, MD

We developed a technique of reverse engineering to extract a conceptual schema — also called "categorial structure" in the European standard CEN ENV 12264 (MoSe) — from a set of terminological phrases. The technique was originally applied to coding systems, ie. to large value sets.
We applied this technique to subsets of two new terminological resources for message standards: headings of patient record from Clinical LOINC and names of "Context Groups" for structured reporting from SNOMED DICOM Microglossary (SDM). Both sources provide context-independent names for message fields and domains of admitted values.
Therefore conceptual schemata on the potential content for a field are compatible with the ones on names of the fields themselves.
Both kinds of schemata can be compared and integrated with conceptual schemata for the information system that manages the patient record.
This continuity in the schemata allows the coupling of applications with different organization of data, and will facilitate mapping from an application to standard messages and viceversa.
Moreover, the simplified representation produced according the MoSe approach is easy to understand by healthcare operators, allowing their progressive involvement in cooperative efforts of design, discussion and validation of the schemata.

1. INTRODUCTION

Developers of information systems perceive coding systems as a separate world; similarly, developers of coding systems are not able to cross the boundary towards the structure of patient records. A computer cannot integrate record items from different patient record systems (or the clinical content of a message), unless the structure is the same in all the systems.
In this paper we identify a continuum in the terminology used in patient records — from content, to field name, to context — and show how they may be handled in a uniform way. Using this approach, we show how appropriate semantic analysis allows the integration of terminologies and patient records.

Name, content and context of a record item
The prestandard CEN ENV 12265 [1] on Electronic Healthcare Record Architecture requires that a 'record item' (having a 'name' and a 'content') is a part of a 'record item complex', that in turn can be part of record item complexes of higher rank. This chain of complexes is the 'context' for the record item.
For example, in a given record system the value:
"myocardial infarction"
could be the content of a heading which has the name:
<circulatory diseases> .
It must be always accompanied by a context, such as:
<family history>, [chapter]
<mother>, [sub-chapter]
which allows the correct reconstruction of the information in the complete record item, ie.:
patient's mother had a myocardial infarction

But the same information could be represented in other ways, according to the strategies of developers to satisfy diverse users' requirements, for example:

1) as a value from a short ad-hoc menu, corresponding to a sub-heading under a detailed heading:
<family history of myocardial infarction>
<family member> = "mother"

2) as a binary answer to a hyper-specific question (perhaps in a multicenter study on a narrow topic):
patient's mother had a myocardial infarction = YES

3) as two values, corresponding to two sub-headings:
<family history>
<circulatory diseases> = "myocardial infarction"
<family member> = "mother"

Note that a value (eg. "*myocardial infarction*" in a field *<condition>*) must be interpreted according to the chain of complexes that makes its context, eg.:
- current problem of the patient;
- risk of problem for the patient;
- past history of the patient;
- past history of a member of patient's family, etc.

In other words, all the three elements of a record item (name, content, context) are needed to reconstruct the complete information, and therefore they must be always transmitted as a whole within messages.

Fragments of information and structure
The various information components, in our case:
<person> = "mother",
<temporal tag>= "past",
<condition> = "myocardial infarction"
can be distributed in many different ways among content, name of the heading, and its context.
To be able to process data, passing from scattered elements to the integrated phrase and vice-versa, we

1091—8280/97/$5.00

should know the combination rules, eg. a fragment of a semantic network at a generic level, called pattern:

<condition>
pertains to <person>
has temporal marker <temporal tag>

and also some hierarchical relations, as:

"myocardial infarction"
is a kind of <cardiovascular disease>
is a kind of <condition>
"mother"
is a kind of <family member>
is a kind of <person>

Such transformations can be performed with the assistance of a Terminology Server, able to manage patterns and hierarchies according to a formal model of semantics in health care, as the TeS developed in the GALEN project [2].

Item names in message standards

New initiatives in message standards, as SDM [3] and Clinical LOINC [4], suggest an original perspective on the issue. Both provide a list of item names (called respectively context groups and headings); for each name, they provide a set of allowed values (called respectively context group table and answer list).

Item names may regard a field in a record or in a message; they are themselves expressions that can be analyzed by our techniques of semantic modelling.

For example, the heading

<family history of disease>

can be interpreted as a semi-instantiated pattern like:

<condition> that
pertains to <family member>
has temporal marker "past"

Analogously, from the heading of Clinical LOINC:

<site of collection of stools>

we can extract the following pattern:

<structure> that
is site of "collection" that
has target "stools"

Note that in this example we made explicit the idea of *<structure>* (in fact, the set of allowed answers includes *<body part>* and *<device>*), and that the idea of *site* is expressed by the semantic link *is-site-of*.

The same heading could evoke also another kind of pattern, making explicit the context of an observation process and rearranging the slots:

"observation process" that
has phase "collection" that
has target "stools"
has site <structure>

Our approach of reverse engineering provides rules to extract, manipulate and integrate hundreds of such patterns into a comprehensive set of frames and hierarchies [5,6]; a few examples are shown in § 3.

The approach can be applied to any expression (heading, chapter, field name or value) involved in patient records and message standards. The result is a conceptual schema, called "categorial structure" of the semantic field [7].

Purpose of the paper

We processed item names from Clinical LOINC and SDM, to obtain an initial version of their categorial structures. We outline in § 4 the potentialities of the approach in the near future, with respect to:

- cooperative production of principled and coherent terminological systems, by iterative virtuose loops of validation, application and refinement;
- production of tools to normalize items of structured reports and patient records with different formats;
- design of dynamic bridges to exchange information between independent patient record systems.

2. MATERIALS AND METHODS

Experimental corpora were the item names from two terminological systems for message standards.

The methodology (§ 2.3) stems from the know-how on cooperative modelling acquired in CEN/TC251/ WG2 [7] and in the GALEN Project [2, 6].

2.1 Clinical LOINC

The LOINC database provides a list of items with the respective codes for data interchange on laboratory properties [13]. We used the first release of a "clinical" extension (Dec, 3 1996), that covers the subjects listed in Table 1 and contains 1492 items.

Table 1. Subjects covered to date in Clinical LOINC

Blood pressure (systolic, diastolic, and mean)
Heart rate (and character of the pulse wave)
Respiratory rate
Critical care measures
Cardiac output, resistance, stroke work, ejection fraction, ...
Body weight
Body height
Body temperature
Circumference of chest, thigh, legs
Intake and output
Major headings of history and physical
Major headings of discharge summary
Major headings in operative note
Electrocardiographic measures

Names in Clinical LOINC are unambiguous and include the context. Names are divided into segments. Allocation of fragments of names into segments could be more systematic, ie. precise rules for systematic naming could be developed (see § 4).

We divided items in three classes:

1. names regarding observations and additional details, for which a previous categorial structure prepared for CEN/TC251/WG5 [9] and the structure of the LOINC database on laboratory properties were available;
2. names of features observed on signals and images;
3. headings specific for the patient records.

2.2. SDM

DICOM is a data-interchange standard for biomedical images and image-related data. Recently, supplements on Visible Light [14] and Structured Reporting [15], were released for discussion.

Supplements do not contain terminology, but refer to an external source — SDM, SNOMED DICOM Microglossary [3] — which maintains terminological tables with actual phrases and codes to describe i) the clinical and operational context of the image acquisition procedure and ii) the report, ie. observer judgements evoked by documented evidence.

Tables are organized into 'context groups' (including context in the name). For each context group, a set of admitted values is referenced (an explicit ad hoc list or an external coding system). Each value belongs to one or more context groups [16, 17].

We used release 1.01 (available via Internet [3[) that contains 33 context groups. Further processing of some additional 200 context groups (not yet released) is in progress. Our intent is to contribute as early as possible to the "production cycle" (see § 4).

2.3 Method

Our approach included the following activities, embedded in cycles of evolutionary refinement:

1. collect field names and the sets of allowed values;
2. transform each name to make the context fully explicit (also by observing the related set of values);
3. make the paraphrase of each name, working out explicit details and producing a uniform style;
4. assign a category to each paraphrase, and extract its pattern;
5. cluster paraphrases according to similar patterns;
6. analyze and refine the patterns, also revising the arrangement of paraphrases into the clusters;
7. harmonize the whole set of patterns into a coherent categorial structure (or with a pre-existing categorial structure), including a suitable hierarchy;
8. generate new systematic names for the headings according to the coherent categorial structure;
9. make results available to domain experts for discussion, validation and refinement of: categorial structure, headings, and sets of admitted values.

We applied informally activities 1-7 to each of the three classes of LOINC item names, and then we applied to context groups of SDM a more systematic process with logging of intermediate results.

After a few iterations and refinements (independently carried out in the above subdomains and corpora), we plan to merge the various categorial structures, on the basis of a formal ontological analysis [12].

3. RESULTS

Structures extracted by reverse engineering from Clinical LOINC are shown in tables 2 to 4.

Table 2. Tentative structure on observations, as derived from an informal analysis of LOINC sources

```
<property>
    is acquired according to <method>
        has acquisition way <source>
        has duration <temporal value>
        is acquired under <challenge>
        has acquisition site <anatomy>, <device>
        is related to <event>
            has spec <before | after | during | at>
        has spec <position of patient>
        has means <device>
    is processed by <formula type>, <score table>
    has target <body component>, <event>, <function>
        has spec < measure-independent circumstance>
        belongs to <body system>, <person>
            has context <supersystem>, <compartment>
```

Table 3. Tentative structure on observations on signals and images, as derived from an informal analysis of LOINC sources

```
<graphic property>
    pertains to <graphic object>
        derived from <acquired object>
            is tracing of <property>
                pertains to <body part>
                    belongs to <system>
where:
<graphic property> is a shape, a slope, an amplitude, ...;
<graphic object> is a wave, a complex, an opacity, ...;
<acquired object> is the signal or the image under study.
```

Table 4. Tentative structure on major headings for patient record, as derived from an informal analysis of LOINC sources

```
"annotation"
    regards <condition>, <event>
        has spec <temporal marker>
        pertains to <body component>, <body system>
            belongs to <person>
    is obtained from <source>
```

From context groups of SDM we extracted frames (for a total of 70 slots) on:

- procedure;
- anatomic structure;
- chemical substance;
- imaging device;
- lesion;

and a corresponding taxonomy of 69 entities.

We created two kinds of frames, and we assigned slots to them according to the following principles:

- structural frames provide the links to organize the semantic field and to build taxonomies (see table 5);
- secondary frames (see table 6) provide the way to build highly specific terminological phrases, by adding further details. Arrangement into frames refers to semantic behaviour, not to organizational or clinical relevance of data. Both kinds of frames can be nested to explode each category at the required detail.

Table 5. Examples of structural frames extracted by processing names of SDM context groups 1-33. Numbers refer to original context groups.

<procedure>
    has means <device>
    has subprocedure <imaging procedure>
    has subprocedure <diagnostic procedure>
        has target <anatomic structure 01b>
        has target <abnormal structure>
<imaging procedure>
    is specified as <photographic image type 33>
        has target <anatomic region or structure 01>
    uses <viewing device>
<angio interventional procedure 09>
    is guided by <imaging device 08>
        employs <chemical substance 10>
<radiographic procedure>
    employs <contrast agent 12>
        has component <active ingredient 13>
<nuclear medicine procedure>
    employs <radiopharmaceutical agent 25>
        has component <diagnostic radionuclide 18>
    according to <imaging projection 26a>
    according to <beam orientation 26b>

Table 6. Examples of secondary frames from SDM. Numbers refer to original context groups.

<procedure>
    has spec <priority 16>
    has spec <administrative scheduling status 17>
    has spec <success or not>
<imaging procedure>
    has spatial condition <position detail>
        has spec <position wrt gravity 19>
            has spec <specific position 20>
            has spec <tilt angle 29>
        has spec <position wrt gantry 21>
        has spec <cranio caudal angulation 23>
        has spec <extremity position 27>
    has approach <approach 24>

4. DISCUSSION

In the discussion we will consider the rationale for the method, the results, and the potential implications.

Involvement of domain experts

The formalism we use is intentionally simple, but is precise at the needed level of approximation.
It allows the involvement of a large number of users after a minimal training. Since LOINC and SDM both encourage proactive participation of medical societies and manifacturers, barriers of formalism or pressure for extreme precision should be avoided.
The structured representation acts as an intensional definition, in the particular vision of world embedded in a structure. It allows comparisons across domains and quality assurance; it facilitates production of rules for systematic names in a multilingual environment.
Experts should master their formalizations and should be brought to an adequate level of understanding of the semantic mechanisms, in order to fix and validate models in an progressive process of refinement.
Cooperation among domain experts requires coherence among different visions of world and reconciliation into a unique structure, avoiding unnecessary diversity but preserving and fixing real differences. More in general, overlaps between independent initiatives can be harmonized, eg. between the items on orders in LOINC and procedures performed in SDM.

Appropriateness of 2nd generation systems

Classifications or nomenclatures are terminological systems of the 1st generation [5]. Their long lists, even if hierarchically organized, cannot cover the new needs of structured reports and patient records.
In fact, health care provision deals with motivated, ad hoc terminological phrases of arbitrary length, according to the following "law":

"given a phrase from a nomenclature with an arbitrary level of granularity, every user's particular task will require an additional detail"

New systems, as SDM and LOINC, are based on a compositional approach and allow user-controlled extensions, according to generic compositional constraints; the user is responsible for creating sensible combinations (cfr. 2nd generation in [5]).
Formal systems (3rd generation in [5]) have a proper engine to generate all and only the sensible combinations. They show exciting performances but require resources to build and validate the underlying model. Moreover, adequate skills are still rare: formal modelling is not familiar to healthcare professionals.
Nevertheless, a second-generation system is able to convey relevant structured knowledge. The GALEN project is successfully refining methods and software tools to perform semi-automatic mapping from an intermediate representation using a second-generation system to a formal model of 3rd generation [6, 11].

Transmission of record items

We are exploring the benefit to patient record when information is interpreted and stored as *structured data*, eg. by means of a structured interface or by a natural language processor. And then it is transmitted from one record system to another via standard messages.
Effective transmission could be obtained in two ways:
- impose a fixed record structure (not realistic);
- develop principles and tools for dynamic coupling.

Continuum among names, content and context of record items is a key issue for network-based patient record and for messages in health care telematics.
Our underlying hypothesis is that developers of patient records, data interchange message standards, and terminologies must share the same very general conceptual schema, in order to understand how to distribute semantic information between
- structural items of their system, ie. field names, data element names, and headings;

- patient-related content of the record (actual patient-related values)

and to rearrange such information into meaningful phrases in context.

Mapping of schemata can be made during design and customization activities, or be dynamically performed by computer as a run-time task. The first solution is limited, but easier to realize by semi-automatic mapping of each existing structure to a generalized format, to initialize and customize "bridge" software.

Methods and software for dynamic normalization of an expression (suitable also for complete record items) were developed in GALEN [2]. They could allow the **reading** of items from different formats.

Methods for dynamic fragmenting a record item into a predefined structure (needed for **querying** and **writing**) are still to be developed.

5. CONCLUSIONS

Item names for message standards and corresponding admitted values were developed by domain experts using a pragmatical approach.

Our technique allows the refinement of the original material to obtain more clear, coherent and unambiguous item names.

There is a limit that can be appreciated and tolerated today by experts and users in the precision of medical expressions. The situation will evolve if there is a force towards systematization, eg. if computer-assisted semantics will provide a perceived benefit.

An evolutionary gradual process of awareness and involvement, with the related cultural shift has to be put in place, involving also the academic community and the formative process of healthcare operators.

Eventually two approaches are converging:

- methods for cooperation and principles for semantic integration in Europe (CEN and GALEN);
- pragmatical activities in US (LOINC and SDM).

Synergy across the Atlantic can now be achieved, with active involvement of experts and users in the preparation and maintenance of terminological resources for message standards.

The new terminology resources will foster evolution of applications towards appropriate management of semantics, based on integration of schemata from terminology, patient record and information system.

Acknowledgments. Funding for this work was provided by the European Union (GALEN and GALEN-IN-USE Projects), the Italian Ministry of Universities and Research (Telemedicine Program, tema 1), the National Library of Medicine.

The Authors wish to thank Clem McDonald, Stan Huff, Alan Rector, Werner Ceusters, David Lloyd, for the productive discussions on the topics of this paper.

References

1. CEN ENV 12265: 1995. Medical Informatics — Electronic Healthcare Record Architecture
2. GALEN and GALEN-IN-USE documentation (1992-96): http://www.cs.man.ac.uk/mig/galen
3. Bidgood WD Jr. ed. SNOMED DICOM Micro-glossary. Coll. Am. Path. Northfield, IL 1997. http://www.snomed.org/sdm/sdm.htm
4. http://www.mcis.duke.edu:80/standards/HL7/termcode/loinclin/loinclin.html
5. Rossi Mori A, Consorti F, Galeazzi E. Standards to support development of terminological systems for healthcare telematics. Meth Inform Med., 1997 (to appear)
6. Galeazzi E, Rossi Mori A, Consorti F, Errera A. A methodology to build conceptual models in medicine. in Proc MIE 97, IOS Press, 1997, pp 280-284
7. CEN ENV 12264: 1995. Medical Informatics — Categorial structure of systems of concepts — Model for representation of semantics
8. Digital Imaging and Communications in Medicine (DICOM). PS 3.1 - PS 3.12. NEMA, VA 1992, 1993, 1995, 1997
9. CEN/TC251 PT5-021 Vital Sign Information Representation, 1996
10. Coté RA, Rothwell DJ, Palotay JL, Beckett RS, Brochu L, eds. The Systematized Nomenclature of Human and Veterinary Medicine. Northfield, IL. College of American Pathologists 1993
11. Rogers JR, Solomon Wd, Rector AL, Pole P, Zanstra P, van der Haring E. Rubrics to dissections to GRAIL to classifications. in Proc MIE 97, IOS Press, 1997, pp 241-245
12. Steve G, Gangemi A. ONIONS Methodology: Ontological Commitment of Medical Ontology. in B Gaines, M Musen (eds), Proc. Knowledge Acquisition Workshop, Univ. of Calgary 1996
13. http://www.mcis.duke.edu:80/standards/termcode/loinc.htm
14. Digital Imaging and Communications in Medicine (DICOM). PS 3 Suppl.15: Visible Light Image for Endoscopy, Microscopy, and Photography. NEMA. Rosslyn, VA 1997
15. Digital Imaging and Communications in Medicine (DICOM). PS 3 Suppl.23: Structured Reporting. NEMA. Rosslyn, VA 1997
16. Bidgood WD Jr. The SNOMED DICOM Microglossary: Controlled Terminology Resource for Data Interchange in Biomedical Imaging. Meth Inform Med., 1997 (to appear)
17. Bidgood WD Jr. Documenting the Information Content of Images. JAMIA Supplement, Proceedings of the Twenty-First Annual Meeting of the American Medical Informatics Association (submitted)